Dose-Response, 6:352–368, 2008
Formerly Nonlinearity in Biology, Toxicology, and Medicine

ISSN: 1559-3258
DOI: 10.2203/dose-response.08-007.Hunt

International Dose-Response Society
www.Dose-Response.org

SUMMARY OF DOSE-RESPONSE MODELING FOR DEVELOPMENTAL TOXICITY STUDIES

Daniel L. Hunt □ Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN

Shesh N. Rai □ Biostatistics Shared Facility, JG Brown Cancer Center, and Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY

Chin-Shang Li □ Division of Biostatistics, MS1C, Room 145, Department of Public Health Sciences, University of California, Davis, CA

□ Developmental toxicity studies are an important area in the field of toxicology. Endpoints measured on fetuses include weight and indicators of death and malformation. Binary indicator measures are typically summed over the litter and a discrete distribution is assumed to model the number of adversely affected fetuses. Additionally, there is noticeable variation in the litter responses within dose groups that should be taken into account when modeling. Finally, the dose-response pattern in these studies exhibits a threshold effect. The threshold dose-response model is the default model for non-carcinogenic risk assessment, according to the USEPA, and is encouraged by the agency for the use in the risk assessment process. Two statistical models are proposed to estimate dose-response pattern of data from the developmental toxicity study: the threshold model and the spline model. The models were applied to two data sets. The advantages and disadvantages of these models, potential other models, and future research possibilities will be summarized.

Keywords: Developmental toxicity study, Dose-group variation, Estimation, Spline, Threshold

INTRODUCTION

Developmental toxicity studies involve the investigation of the responses of fetal litters to maternal exposure to a potentially toxic agent, which while it may cause negative effects, is considered to be a non-carcinogen for humans. The U.S. Environmental Protection Agency (USEPA) is the primary federal protection agency that uses the results of these studies to develop guidelines for safe levels of human exposure to these toxic agents, which can manifest themselves naturally in the air, but also in common man-made products and structures that humans are exposed to in the home, workplace, and the general public. Therefore, USEPA works in collaboration with other regulatory agencies such as the

Address correspondence to Daniel L. Hunt, Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA 38105-2794, USA; Tel: (901) 495-5501; Fax: (901) 544-8843; Email: daniel.hunt@stjude.org

Food and Drug Administration (FDA) and Occupational Safety and Health Administration (OSHA) to establish guidelines. Additionally, one or more of the most recent USEPA guidelines for areas of risk assessment associated with developmental toxicity, may be combined with (or may supersede) the USEPA developmental toxicity guidelines when exposure is under suspicion of resulting in additional toxicity not accounted for in the developmental guidelines (USEPA, 1986; 1992;1996; 1998; 2005).

In the developmental toxicity study, each impregnated animal is randomly assigned to a dose group corresponding to an exposure level of the toxic agent under study. The agent is usually administered orally to the animals during fetal development. Endpoints are then measured and recorded on both the animals and their fetal litters. A dose-response relationship that relates agent dose level to these endpoints is assumed to exist. Some method for determining tolerable level of exposure is used. This should primarily involve estimating the dose-response pattern. At the conclusion of the risk assessment process when a dose-response relationship has been estimated, the results are then extrapolated to determine safe exposure levels of the toxic agent in humans during fetal development (USEPA, 1991). USEPA considers statistical modeling to be an important step in the risk assessment process (Ryan, 2000).

The major endpoints that are measured on fetuses are deaths, structural malformations, growth aberrations, and functional deficiencies. Also, endpoints relating to fetal weight and length are measured and modeled using some continuous distribution. Data relating to endpoints such as indicators of death and malformation are categorical and these are the endpoints that we investigate in this paper. An endpoint of death or malformation is categorized as an 'adverse event', with death superseding malformation for a given fetus. Since each litter is a natural cluster, individual litter measures are typically accumulated into a sum or average and that quantity is seen as a data value to represent the entire litter, e.g., number of deaths or malformations in the litter, or average litter weight. Thus, the study sample size equates to the total number of litters. Even after the grouping, the sample size is still adequately large enough for the study to have appreciable power and for the estimates to be meaningful. This is not always the case as data can be left as individual-level and modeled accordingly. Typically, individual-level data is used in cases of jointly modeling bivariate outcomes, such as fetal weight and malformation (Catalano and Ryan, 1992).

Developmental toxicity studies typically involve investigation of environmentally unsafe agents, which, when exposed to an unacceptable level, results in non-carcinogenic toxic effects. Since carcinogenic risk assessment studies are restricted to use of the linear dose-response model, i.e., the linear-no-threshold (LNT) model, as the default model (USEPA, 2005), the developmental toxicity study is not tied down by such limita-

tions. In general, the default model in the assessment of non-carcinogenic risk is the threshold dose-response model (USEPA, 1991). In the context of the typical single-agent and endpoint dose-response study, the *threshold* is defined as the maximum dose level at which the toxic response equates to the background control level response. Threshold is assumed to exist for developmentally toxic agents due to biological ability of organism to defend itself against tolerable level of toxic threat.

Although the USEPA inherently considers threshold to be the default for studies assessing non-carcinogenic risk, they do not currently employ pure threshold modeling techniques in assessing risk, i.e., they do not obtain estimates based on the pure threshold model. Instead, they use two approaches: (1) the no-observed-adverse-effects-level (NOAEL) approach and (2) benchmark dosing. The NOAEL approach simply identifies, among the dose groups under study, the highest experimental dose group which is not statistically significantly different from the control level in terms of response. In this way, the threshold could be assumed to exist between the NOAEL and the next lower dose level from the NOAEL. The NOAEL approach does not use a parametric model, however, the benchmark dose approach does. Introduced by Crump (1984), the benchmark dose is the lower confidence limit of a dose level, estimated by a parametric model, which yields an acceptable level of excess risk (above control level risk). In both cases of NOAEL and benchmark dosing, a safety factor is applied to the final dose and this is determined as the acceptable level of exposure for the human population.

The most obvious advantage of using benchmark dosing over NOAEL is that benchmark dosing presumes a dose-response relationship, and hence one can obtain estimates of response as well as of the benchmark dose itself. The assumed model is highly data-dependent, i.e., the proposed model should accurately reflect the pattern of the data. This is true of any parametric model, including the threshold model. Another advantage is that, being a parametric model and necessitating estimation of parameters, the benchmark dosing approach takes into account the response variation within each dose group, also an important factor in estimating risk. Although benchmark dosing estimates a tolerance level of sorts, it still does not directly estimate the threshold, not an easy undertaking. The advantage of the threshold model is that it directly estimates the point of change in the direction of the responses in the dose-response curve.

Figure 1 is a typical threshold dose-response model. It represents the expected threshold dose-response curve for effects that increase with increasing levels of toxic exposure, such as incidence of death or malformation. This is similar to the set of threshold dose-response models described by Cox (1987), who examined several data sets extracted from general toxicology studies involving insects and animals. While the exper-

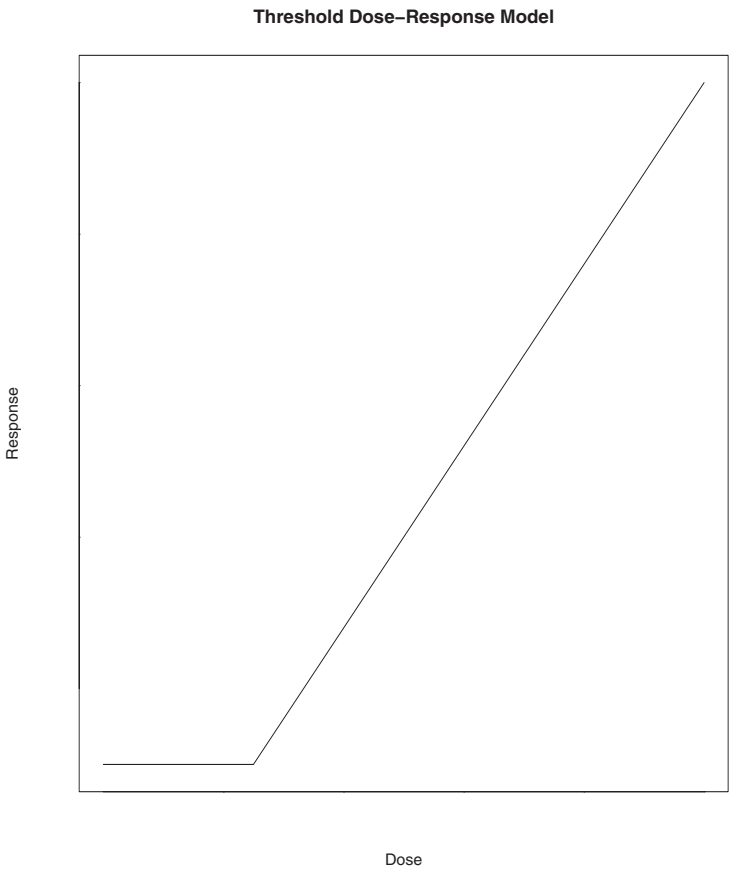

FIGURE 1. Threshold dose-response model; response increases with dose above threshold.

imental units were directly exposed animals and insects, the outcome of interest was mortality rate, resulting in a dose-response pattern similar to Figure 1. Also, he used likelihood-based approaches to derive estimates, as we do in this paper.

The beta-binomial distribution was first proposed by Williams (1975) to model the number of adverse fetal responses. It has the desirable quality of having a parameter, the intralitter correlation, to account for litter effects. Chen and Kodell (1989) first used the beta-binomial in the context of dose-response modeling by fitting beta-binomial with a monotonic dose-response model to a developmental toxicity data set. Hunt and Bowman (2004) fit a hormetic model to that same data set, with beta-binomial distribution. A limitation of the beta-binomial approach to dose-response modeling is that the correlation parameter is modeled separately from the dose-response parameters linked to the parameter equating to response. Hunt and Rai (2003) introduced a desirable alternative to the beta-binomial. They assumed a binomial distribution for the litter responses conditional on litter response variation being from a normal

distribution. The advantage is that the parameter for response variation is part of the dose-response model, thereby facilitating estimation.

Hunt and Rai (2003) modeled response variation as being uniform across dose groups, primarily to limit the nuisance parameters to be able to compare bias in estimating threshold between their model and the beta-binomial. Kupper *et al.* (1986) simulated monotonic dose-response models with the beta-binomial distribution and found that assuming a single correlation parameter can lead to estimation bias if the underlying model is one of multiple correlation across dose groups. However, they used very limited number of dose groups (three) and this alone could contribute to the bias. Hunt and Rai (2008) added multiple variation parameters to their model to test this assumption with more dose groups (five, more typical of what is tested in these studies) and, while finding bias in the single-variation model, also found that the bias was still very low for cases where there was relative closeness in the variation across dose groups.

Li and Hunt (2004) introduced a regression spline approach which counters the threshold approach by having the ability to model several directional-changing dose-response patterns, including the threshold model itself. Similar to Hunt and Rai (2003), they assumed uniform response variation. As Hunt and Rai (2008) illustrated that the multiple-variation model was significant over the single-variation model for their investigated data set and their simulations showed that the single-variation model can lead to bias, we use the multiple-variation model here. We describe the models in the Methods section. In the Results section, we compare and contrast the threshold and spline modeling approaches to data sets from two developmental toxicity studies. The Discussion concludes the paper by summarizing the comparative merits of the proposed models and discussing the potential for future work.

MATERIALS AND METHODS

Environmental Agents

We investigated the applicability of our model to data extracted from two developmental toxicity studies. The first study involved the investigation of the plasticizing agent diethylhexyl phthalate (DEHP), which is a potential hepatic carcinogen for rodents (Doull *et al.*, 1999). The second study investigated animal exposure to the organic solvent diethylene glycol dimethyl ether (DYME). Both agents are in the class of agents that are suspected developmental toxicants for humans (USEPA, 1991).

Endpoints

For both studies, upon animal sacrifice, dams were evaluated for several endpoints, including the number of implantation sites and the fetal

TABLE 1A. Summary of data by dose level from the DEHP study

Level of DEHP (%diet)	Number of dams	Average litter size	$\bar{P}$ ± SE	Range of P
0	30	13.2	0.188 ± 0.206	0-0.625
0.025	26	12.3	0.121 ± 0.099	0-0.0375
0.05	26	12.3	0.253 ± 0.183	0-0.643
0.10	24	11.5	0.723 ± 0.313	0.143-1.000
0.15	25	12.3	0.982 ± 0.078	0.615-1.000

TABLE 1B. Summary of data by dose level from the DYME study

Level of DYME (mg/kg/day)	Number of dams	Average litter size	$\bar{P}$ ± *SE	Range of P
0	21	14.1	0.059 ± 0.070	0-0.025
62.5	20	12.1	0.095 ± 0.096	0-0.333
125	24	13.0	0.110 ± 0.093	0-0.308
250	23	13.0	0.339 ± 0.258	0-0.846
500	23	12.4	0.973 ± 0.082	0.615-1.000

status of each site. Fetal status is categorized by resorbed, dead, or live fetus. Live fetuses were further investigated for body weight and structural, visceral, and skeletal abnormalities. For the purposes of our analysis, with each animal being a unit, the fetal litter of an animal was grouped together with a fetal resorption, death, or abnormality counting as an adverse event and any other outcome as a non-adverse event.

Animal Studies

In the first study, 131 timed-pregnant CD-1 mice were randomly allocated to be exposed to one of five levels (0.0%, 0.025%, 0.05%, 0.10%, and 0.15%) of DEHP in their feed on gestational days 0 to 17 (Tyl *et al.*, 1983). In the second study, 111 timed-pregnant CD-1 mice were allocated to be orally administered one of five levels of DYME (0, 62.5, 125, 250, or 500 mg/kg/day) on gestational days 6 to 15 (Price *et al.*, 1987). For both studies, animals were sacrificed at day 17 for evaluation of all dams and their fetuses. Summarized results of both studies is given in Table 1A (DEHP study) and Table 1B (DYME study).

From Table 1A, the results of the DEHP study, it can be seen that the average proportion of responses changes dramatically in the low-to-intermediate dose range (0-91 mg/kg/day). There is sharp decrease in response from the control level (about 19%) to the response at the second dose level 44 mg/kg/day (about 12%), followed by the expected increase thereafter; this pattern seems to imply that there could indeed be a threshold dose level for this study. For Table 1B, the DYME study, the pattern isn't quite as conclusive. However, the response at the second and

third dose levels 62.5 and 125 mg/kg/day (9.5 and 11%, respectively) remain relatively close to response at the control level (about 6%). Additionally, the SE estimates indicate extreme overlap in the responses at the three lowest dose levels.

As noted in the Endpoints section, the outcome to be modeled is the number of adversely affected fetuses in each fetal litter. For a given litter, let represent this number. Subsequently, let n be the total number of fetal implants, which equates to litter size for the purposes of our analysis. Then the response in each litter is represented by the proportion of adversely affected fetuses, i.e., $P = X/n$. Since we are dealing with discrete binomial endpoint, a logistic link function for the dose-response model is applicable. Also, the dose-response function must be piecewise to account for the threshold dose. Additionally, as litters are expected to have response variability, overdispersion should be a factor in the model.

The Threshold Dose-Response Model

Define $P(d)$ to be the probability of toxic response at dose level d. Here, $P(d)$ is the probability of adverse litter effects. Because the data here is discrete and binary, the logistic link function is used (McCullagh and Nelder, 1983), hence the threshold model with multiple response variation is given by:

$$\text{logit}[P(d_i)] = \theta_0 + \theta_1(d_i - \tau) \times I(d_i > t) + \sigma_i Z, \qquad (1)$$

$$i = 1, \ldots, g$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \tau, \sigma_1, \ldots, \sigma_g)$ is the parameter vector, with τ being the threshold dose level, σ_1^2 is the random response variation for in the ith dose group d_i and I is the indicator function for the dose being higher than the threshold.

The Spline Approach

Li and Hunt (2004) introduced an approach that uses linear B-splines to replace the regular covariates in the standard dose-response function, but still incorporates the variation parameter as first introduced by Hunt and Rai (2003). The regression spline approach includes an interior knot which can be interpreted as a changepoint, but not necessarily the threshold, in the direction of the dose-response pattern. It has the desirable aspect of being able to model several dose-response patterns that could occur in the data, including the pure threshold model. One could include more knots or assume higher degree splines (quadratic, cubic), but these assumptions necessitate the addition of more parameters, thereby requiring a sufficient amount of data to still have desirable study power.

The general form of the polynomial regression *B*-spline function is given by the following equation:

$$s(d; \boldsymbol{\theta}, \boldsymbol{\varepsilon}) = \sum_{j=1}^{m+k} \theta_j B_j(d; \boldsymbol{\varepsilon}), \tag{2}$$

where m is called the polynomial order (one more than the degree) and k is the total number of interior knots, i.e., $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)'$; the set $\{B_i(d; \boldsymbol{\varepsilon}): j = 1, \ldots, m + k\}$ is the set of B-splines of order m (degree $m - 1$) constructed recursively starting with a set of order 1 B-splines, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{m+k})'$ is the set of B-spline coefficients. This theory is described in de Boor (2001).

Li and Hunt (2004) use the general function given by equation (3) by setting m=2 (for degree 1 polynomial, or linear) and k=1 interior knot. Since we assume the multiple-variation model of Hunt and Rai (2008) to be the default model, Li and Hunt's model is altered appropriately to allow for multiple response variation to make adequate comparison. Hence, the general form of the new regression *B*-spline function is given by the following:

$$\text{logit}[P(d_i)] = \sum_{j=1}^{3} \theta_j B_j(d_i; \varepsilon) + \sigma_i Z, \tag{3}$$

where all parameters are as described previously in equations (1) and (2).

Nested Models

From de Boor (2001), the spline component of equation (2) can be re-expressed in the following alternate form:

$$s(d; \boldsymbol{\theta}, \varepsilon) = \theta_1 + \theta_2 d + \theta_3 (d - \varepsilon)_+, \tag{4}$$

known as the truncated power basis form. From equation (4), if $\theta_2 = 0$, then it becomes:

$$s_1(d; \boldsymbol{\theta}, \varepsilon) = \theta_1 + \theta_3 (d - \varepsilon)_+, \tag{5}$$

which is equivalent to the linear predictor of the threshold model given by equation (1). That is, with s_1 corresponding to some appropriate transformation of the data, such as the logit, θ_1 is the background response occurring from control dose level $d = 0$ to $d = \varepsilon$. Here, in equation (5), the interior knot ε actually equates to threshold. Hence, the threshold model is a subset of the spline model.

TABLE 2A. Estimates from fitting threshold model to DEHP data.

Parameter	Estimates ± SE
β_0	–2.037 ± 0.170
β_1	60.949 ± 6.895
τ	0.038 ± 0.006
σ_1	1.219 ± 0.220
σ_2	0.402 ± 0.310*
σ_3	0.959 ± 0.276
σ_4	1.851 ± 0.348
σ_5	1.100 ± 0.339

*: not statistically significant based on partial t-testing.

TABLE 2B. Estimates from fitting spline model to DEHP data.

Parameter	Estimates ± SE
θ_1	–2.006 ± 0.352
θ_2	–2.108 ± 0.277
θ_3	5.519 ± 0.831
ε	0.036 ± 0.006
σ_1	1.426 ± 0.266
σ_2	0.009 ± 0.823*
σ_3	0.783 ± 0.223
σ_4	2.554 ± 0.770
σ_5	1.947 ± 0.472

*: not statistically significant based on partial t-testing.

RESULTS

DEHP study

The results (estimates and SEs) from fitting models (1) and (3) to the DEHP data are given in Table 2. For the threshold model, the estimate of τ is about 0.038% DEHP (Table 2A), between the first two experimental dose levels 0.025 and 0.05%. Partial t-testing based on the SE estimate confirms this location as well. This result agrees with the NOAEL. The results do indicate that there is strong evidence for the existence of a threshold dose, indicating that at the very least, a pure threshold model would be applicable. The LRT for threshold significance results in p-value= 0.0001, giving firm support to the existence of a threshold dose level. The LRT results for the test of multiple-variation resulted in a p-value=0.001, indicative of differing degrees of response variation across dose groups. For the spline model, the estimate of the interior knot ε is about 0.036% (Table 2B). Partial t-testing based on the SE estimate confirms this location as well. The LRT results for testing the significance of the spline model (above the threshold model) yields a p-value of 0.012. Hence, there is reason to believe that the spline model is the more appro-

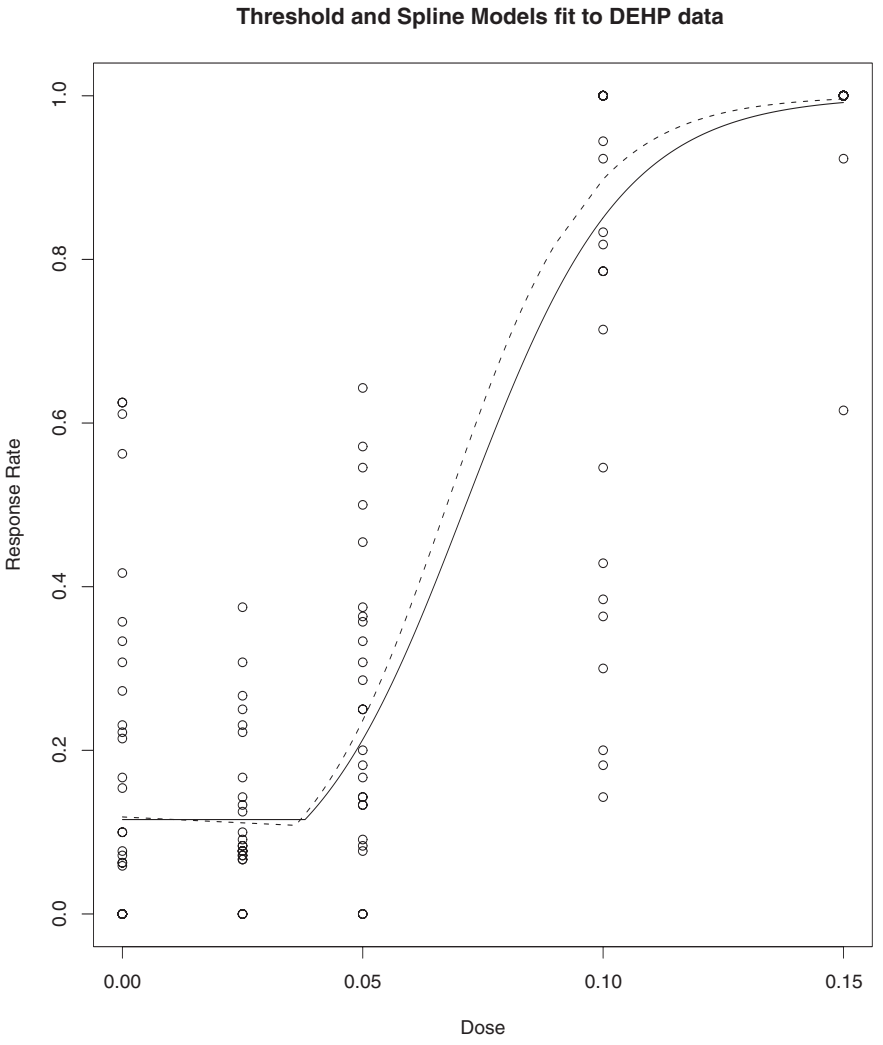

FIGURE 2. Estimated dose-response curves to DEHP data (circles); solid line is the predicted threshold curve; dashed line is the predicted spline curve.

priate model for this data. Figure 2 is the estimated dose-response curve of both models fit to the DEHP data.

The observed pattern of the data in Figure 2 suggests a decreasing dose-response relationship at the lowest dose groups, then expected increase thereafter. Noticeably, from Table 1A, the background response is high at around 19%; then first dose group above control has lower response rate of roughly 12%, with monotonically increasing response after this dose group. This could be due to an unduly high background rate, but the results suggest that at the least, a threshold dose level does exist. The spline model fits a decreasing dose-response function below

TABLE 3A. Estimates from fitting threshold model to DYME data.

Parameter	Estimates ± SE
β_0	–2.641 ± 0.195
β_1	0.022 ± 0.006
τ	154.5 ± 41.3
σ_1	0.309 ± 0.599*
σ_2	0.687 ± 0.384*
σ_3	0.820 ± 0.308
σ_4	1.158 ± 0.263
σ_5	1.855 ± 0.702

*: not statistically significant based on partial t-testing.

TABLE 3B. Estimates from fitting spline model to DYME data.

Parameter	Estimates ± SE
θ_1	–2.898 ± 0.302
θ_2	–1.457 ± 0.632
θ_3	6.327 ± 1.836
ε	228.2 ± 27.0
σ_1	0.530 ± 0.467*
σ_2	0.468 ± 0.409*
σ_3	0.326 ± 0.406*
σ_4	1.192 ± 0.261
σ_5	3.263 ± 1.275

*: not statistically significant based on partial t-testing.

the interior knot estimated to be at 0.036%, then increasing function above, indicating threshold level to exist somewhere above the estimated knot, which in that sense agrees with the threshold model; but based on the estimated background response rate of the spline model (which is slightly above the rate estimated by the threshold model), the threshold estimate of the spline model, 0.041%, is somewhat above that of the one from the threshold model. Above threshold, the two models closely overlap in terms of response estimates.

DYME study

The results (estimates and SEs) from fitting models (1) and (3) to the DYME data are given in Table 3. The threshold model estimate is 154.5 mg/kg/day (Table 3A), between the 3rd and 4th dose levels 125 and 250 mg/kg/day. Partial t-tests indicates estimate is between these levels with 95% confidence. The threshold result also agrees with the NOAEL. The LRT for threshold significance results in p-value= 0.001, indicating that threshold dose level may exist for this data set. The LRT results for the

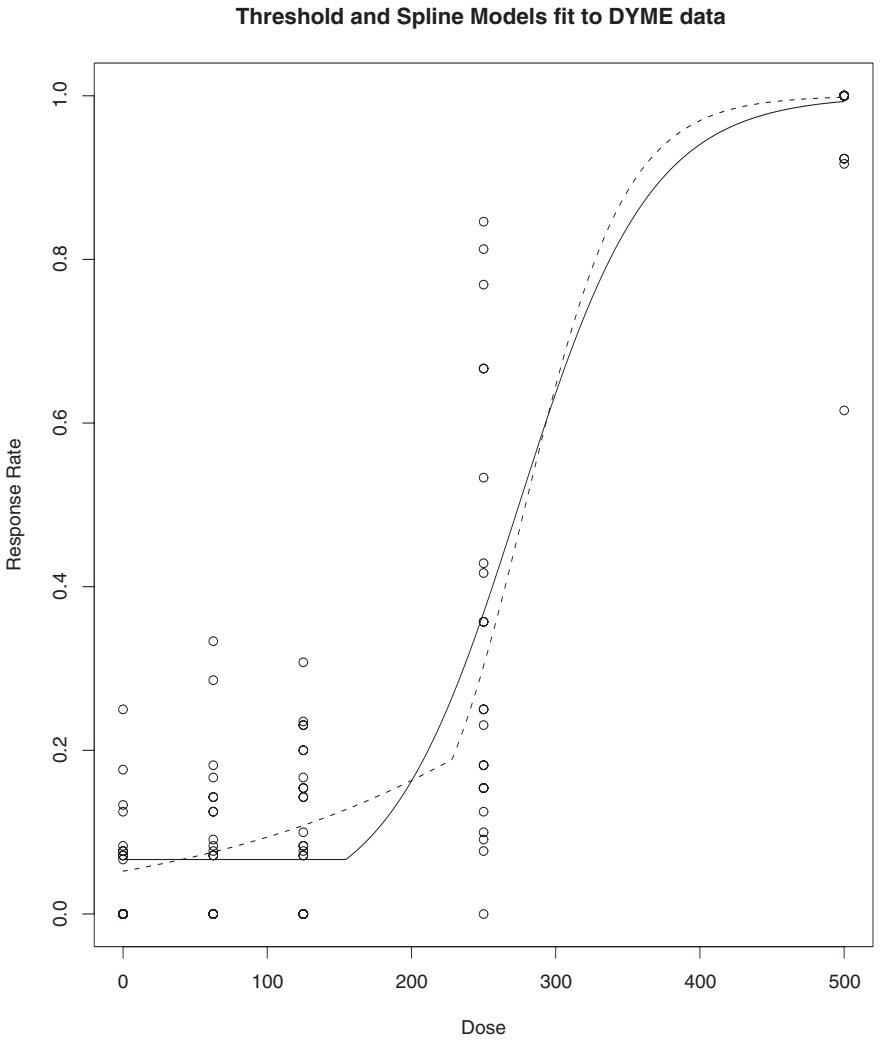

FIGURE 3. Estimated dose-response curves to DYME data (circles); solid line is the predicted threshold curve; dashed line is the predicted spline curve.

test of multiple-variation resulted in a p-value=0.156, which indicates that the single-variation model may suffice for this data set. For the spline model, the estimate of the interior knot ε is about 228.2 mg/kg/day (Table 3B). The LRT results for testing the significance of the spline model yields a p-value of 0.013, indicating that the spline model may be the more appropriate model for this data set. Figure 3 is the estimated dose-response curve of both models fit to the DYME data.

For this study, the threshold estimate 154.5 mg/kg/day is in between the 3rd and 4th dose groups, apparently due to the proximity in the responses (approximately 6%, 8%, 11%) at the lowest three dose groups (0, 62.5, and 125 mg/kg/day, respectively). Also, from Table 1B, the

response variation at these dose groups is very similar, yet the responses technically increase monotonically. Unlike for the previous study, the estimated threshold and knot for this study are not very close. While estimated threshold is 154.5 mg/kg/day, the estimated knot is 228.2 mg/kg/day. The spline curve (dashed) in Figure 3 indicates no threshold as the entire curve is monotonic, however slowly increasing below the knot. Above threshold, both models yield similar response rates, as in the previous study.

Hormetic Effects

The spline curve in Figure 2 implies that hormetic effects may exist in addition to threshold effects. Hunt and Bowman (2004) assumed hormetic effects to exist in the DEHP data, but they assumed the effects to exist in the form of a U-shape curve below the threshold dose level. Their model included a parameter to account for this effect and inherently assumed existence of threshold. While the results showed the hormetic parameter not to be significant, the p-value was much closer to the nominal 0.05 cutoff for significance than the spline model p-value here. Their claim was that this insignificance could very well be likely due to an underpowered study with not enough dose groups as well as not enough litters within dose groups. They investigated this claim in a simulation study and illustrated that it did not have enough power to adequately detect hormetic effects. This lack of power is noticeable from Figure 2 as the difference between the two models is small, while there are too few dose groups below threshold to detect a difference if there is one.

Several authors have noted the low power for detecting hormesis in developmental studies and have laid down some guidelines that should be adhered to when wanting to statistically show hormesis (Sielken and Stevenson, 1998; Teeguarden *et al.*, 2000). Hunt (2002) took this one step further and conducted an extensive simulation study to investigate several factors in the design of developmental studies important in the detection of hormetic effects. He found that the power of hormetic detection can be increased without necessarily increasing the number of dose groups by reallocating the within-group sample sizes (but maintaining the overall size) so that the lower groups have larger sizes and spacing the dose groups properly. Although with the increased power still being low, increasing the number of dose levels, in conjunction with these other factors, should be considered to be able to have probability of hormetic detection.

Doull *et al.* (1999) indicated that there was biological evidence to support the existence of threshold effects in studies of DEHP. Hence, there is some credence to the possibility of its existence in the data. The linear spline model is very close to the threshold model itself, even though it contains the model as a subset. Essentially, it specifically models the changing pattern of the data below and above the knot, which is not far from the

threshold value itself. The hormetic model fit by Hunt and Bowman (2004) was U-shaped and is much different from either the threshold or spline models. These facts, along with the fact that the DEHP study is sufficiently underpowered to detect hormesis, as illustrated by Hunt and Bowman (2004), may indicate that significance of threshold and hormetic effects can be found with more dose groups investigated.

DISCUSSION

We have discussed several threshold dose-response models for application to dose-response data from developmental toxicity studies. There have been various models used, similar in some areas, different in others. In general, the standard form of the pure threshold dose-response function has been used in most cases. The models begin to differ in some of the underlying assumptions about the data itself, e.g., some have assumed the beta-binomial (Kupper *et al.*, 1986; Chen and Kodell, 1989; Hunt and Bowman, 2004), while others have made other assumptions about the data. Schwartz *et al.* (1995) assumed binomial data with simple overdispersion. We assume binomial distribution conditional on random litter effects, as in Hunt and Rai (2003) and Li and Hunt (2004), but with the additional assumption of multiple response variation across dose groups. This assumption has the utility of directly including the effects parameter into the dose-response model, allowing for simultaneous estimation with the dose-response parameters, including threshold.

Certain factors that should be considered when using the model given by equation (3). One is the number of interior knots k. Technically, k can be any number, but for the current design of developmental studies, k=1 seems to suffice. However, if k>1, this immediately adds complexity to the model as now multiple knots must be estimated by the algorithm used and just adding one extra knot, i.e., k=2, can complicate the estimation process. Another factor is the order m of the B-splines. Increasing the order, as well as the number of knots, adds parameters to the model. Hence, one must consider if there is enough data to compensate for these extra parameters. The current design of developmental studies leans toward small values of both m and k, which is very desirable to the estimation process.

Whichever method is used or assumed, the method itself inherently accounts for the litter effects in some fashion, which can be an important consideration. Litter effects have typically been assumed to just be nuisance effects, but proper estimation of these effects can have bearing on the bias in the estimates of the primary model parameters, including threshold. Haseman and Kupper (1979) extensively discussed various models that can be used for data from developmental studies and stressed the importance of using models that accounted for litter effects and

noted that models which ignore litter effects may produce considerable estimation bias in the parameters. Paul (1982) investigated the beta-binomial distribution along with several other versions of the binomial distribution for application to developmental data and found the beta-binomial to perform better in many cases in terms of estimation over the other binomial models. Chen and Kodell (1989), through LR testing, found significance of the beta-binomial distribution fit to the data over the binomial distribution, which ignores intralitter correlation.

While the beta-binomial distribution has been a commonly used distribution for data estimation from developmental studies, its primary use had been in the context of treatment comparison via direct estimation of the mean responses across dose groups (Williams, 1975; Haseman and Kupper, 1979; Paul, 1982). It had been expanded to use in dose-response modeling as well (Kupper *et al.*, 1986; Chen and Kodell, 1989). Yet assuming this model for dose-response can lead to estimation difficulties, especially since the intralitter correlation parameter(s) are separate from the dose-response function in the likelihood. Hence, the alternate model introduced by Hunt and Rai (2003) addresses this issue by directly including parameter into dose-response function. Additionally, through simulation studies, they showed that this alternate model and the beta-binomial model yield similar estimation and bias results when each model is fit to data generated from the other model. Therefore, the model of Hunt and Rai (2003) serves as a suitable alternative to the beta-binomial.

Another consideration that may need to be addressed, in addition to accounting for litter effects, is in accurately accounting for the diversity in effects across dose groups. Kupper *et al.* (1986) conducted simulation studies to address this issue. For beta-binomial assumed data, they compared the model with constant intralitter correlation to one with different correlation across dose groups and found that assuming the constant correlation model can introduce bias in the estimates of the dose-response parameters when the underlying model is one of differing correlations across groups. However, they only had 3 dose groups in their simulations. Hunt and Rai (2008) investigated further and in their simulations, they assumed existence of threshold in the dose-response, used 5 dose groups in each case, and assumed the random litter effects model instead of the beta-binomial; results between these two models are comparable from Hunt and Rai (2003).

From Hunt and Rai (2008), although the general conclusion was the same as in Kupper *et al.* (1986), which is the assumption that uniform dose group variation can lead to bias in parameter estimation, they also found that this bias occurs when there is a large degree of variation across dose groups. That is, when the degree of variation or correlation across groups is small, the model of constant effects can still yield reliable estimates. Assuming different effects across groups necessarily increases the

number of parameters in the model and as the effects parameters are generally seen as nuisance parameters, it becomes necessary to determine if these additional parameters contribute to the model in terms of reducing bias and leading to statistical significance in results. Assuming additional parameters means one extra parameter per dose group, which is not as much of a problem in current settings as the number of groups is very limited in these studies. However, for general usage, it may be incumbent to investigate actual models for the effects themselves which can simultaneously accurately estimate effects and limit bias.

Finally, the comparison of the threshold dose-response model to the linear B-spline model yielded quite similar results. The advantage of the spline approach is in its ability to more accurately estimate the changing dose-response pattern across the dose range. As opposed to the threshold model, the spline model instead estimates the interior knot, which does not necessarily equate to threshold. Other considerations that come into play when assuming spline approach are assuming higher order spline and more knots. For developmental studies as constructed and designed currently, these considerations would seem highly unlikely as there are very limited number of dose levels and therefore limitation in the change of pattern across the range of doses. Yet if the design of these studies do move into this direction, then these considerations might become necessary.

Testing showed that the spline model was not significantly better than the threshold model, but it was noted that this could very well be due to an underpowered study. Currently, as the threshold model is the default model for data from these studies, so it appears to suffice. Even current approaches do not employ the pure threshold model as described here, but use confidence limits and safety factors. That is, the current practice appears to be most focused on determining acceptable levels of exposure rather than direct estimation of any threshold effects. In particular, if the experimental design of developmental toxicity studies is modified to allocate more dose groups, and thereby larger sample size, then the models described here will become even more applicable. There will be a greater ability to more accurately estimate threshold effects.

ACKNOWLEDGMENTS

This publication was partially supported by Cancer Center Support CA21765 from the National Institutes of Health, USA, and the American Lebanese Syrian Associated Charities (ALSAC) (D.L. Hunt). This publication was also supported by University of Louisville Tobacco Excise Tax Funds (GN070315) (S.N. Rai). Finally, this publication was made possible by Grant Number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the

official view of NCRR or NIH. Information on Re-engineering the Clinical Research Enterprise can be obtained from http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp (C.S. Li).

REFERENCES

Catalano PJ and Ryan LM. 1992. Bivariate latent variable models for clustered discrete and continuous outcomes. J Am Stat Assoc 87: 651-658

Chen JJ and Kodell RL. 1989. Quantitative risk assessment for teratological effects. J Am Stat Assoc 84: 966-971

Cox C. 1987. Threshold dose-response models in toxicology. Biometrics 43: 511-523

Crump KS. 1984. A new method for determining allowable daily intakes. Fundam Appl Toxicol 4: 854-871

De Boor C. 2001. A Practical Guide to Splines (revised edition). Sprnger-Verlag, New York

Doull J, Cattley R, Elcombe C, Lake BG, Swenberg J, Wilkinson C, Williams G, and van Gemert M. 1999. A cancer risk assessment of di(2-ethylhexyl)phthalate: application of the new USEPA risk assessment guidelines. Regul Toxicol Pharmacol 29: 27-357

Haseman JK and Kupper LL. 1979. Analysis of dichotomous response data from certain toxicological experiments. Biometrics 35: 281-293

Hunt D. 2002. Dose and litter allocations in the design of teratological studies for detecting hormesis. Teratology 66: 309-314

Hunt D and Rai SN. 2003. A threshold dose-response model with random effects in teratological experiments. Commun Stat Theory Meth 32: 1439-1457

Hunt DL and Bowman D. 2004. A parametric model for detecting hormetic effects in developmental toxicity studies. Risk Anal 24: 65-72

Hunt DL and Rai SN. 2008. Interlitter response variability in a threshold dose-response model. Commun Stat Theory Meth 37: 2304-2314

Kupper LL, Portier C, Hogan MD, and Yamamoto E. 1986. The impact of litter effects on dose-response modeling in teratology. Biometrics 42: 85-98

Li CS and Hunt D. 2004. Regression splines for threshold selection with application to a random-effects logistic dose-response model. Comput Stat Data Anal 46: 1-9

McCullagh P and Nelder JA. 1983. Generalized Linear Models (second edition). Chapman and Hall: London

Paul SR. 1982. Analysis of proportions of affected fetuses in teratological experiments. Biometrics 38: 361-370

Price CJ, Kimmel CA, George JD, and Marr MC. 1987. The developmental toxicity of diethylene glycol dimethyl ether in mice. Fundam Appl Toxicol 81: 113-127

Ryan LM. 2000. Statistical issues in toxicology. J Am Stat Assoc 95: 304-308

Schwartz PF, Gennings C, and Chinchilli VM. 1995. Threshold models for combination data from reproductive and developmental experiments. J Am Stat Assoc 90: 862-870

Sielken RL and Stevenson DE. 1998. Some implications for quantitative risk assessment if hormesis exists. Hum ExpToxicol 17: 259-262

Teeguarden JG, Dragan Y, and Pitot HC. 2000. Hazard assessment of chemical carcinogens: the impact of hormesis. J Appl Toxicol 20: 113-120

Tyl RW, Jones-Price C, Marr MC, and Kimmel CA. 1983. Teratological evaluation of diethylhexyl phthalate (CAS No. 117-81-7) in CD-1 mice. Final Study Report for NCTR/NTP Contract NO. 222-80-2031 9(c). NTIS NO PB85105674. National Technical Information Service, Springfield, VA

USEPA. 1986. Guidelines for mutagenicity risk assessment. Federal Register 51: 34006-34012

USEPA. 1991. Guidelines for developmental toxicity risk assessment. Federal Register 56: 63798-63826

USEPA. 1992. Guidelines for exposure assessment. Federal Register 57: 22888-22938

USEPA. 1996. Guidelines reproductive toxicity risk assessment. Federal Register 61: 56274-56322

USEPA. 1998. Guidelines neurotoxicity risk assessment. Federal Register 63: 26926-26954

USEPA. 2005. Guidelines carcinogen risk assessment. Federal Register 70: 17765-17817

Williams DA. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31: 949-952